Guardrails at the Edge of the Abyss

A Paradigm Critique of Daniel Kokotajlo's "AI 2027" from the Perspective of Logographic AI

Daniel Kokotajlo, a former core member of OpenAI's alignment team, recently released the report AI 2027 [1], shaking the industry with the stark warning that there is a 70% probability AI will destroy humanity. The report is suffocating to read—it paints a doomsday picture of "superintelligence on the verge of spiraling out of control" through precise risk modeling, sincere fear-mongering, and detailed policy blueprints.

Yet, from the perspective of Logographic AI (表意 AI) theory [2], all of Daniel's efforts are, in essence, installing sturdier guardrails at the edge of an abyss—and he has never asked a far more fundamental question: **Can we build in a different place, so that the abyss itself ceases to exist?**

To understand why Daniel's "guardrail" mindset is a paradigmatic limitation, one must first clarify the fundamental divergence between Logographic AI and mainstream AI. Logographic AI is not another version of existing AI technology; it is a novel paradigm that takes the "Morpho-Root" (形根) as its cognitive primitive, standing in fundamental contrast to the mainstream "phonographic AI" (表音 AI)—namely, the Token-based Transformer architecture [2]. The core divergence is this: Token strips symbols from meaning, forcing AI to "guess the next word" in statistical space; the Morpho-Root, by contrast, insists that every cognitive primitive must carry its topological position within the network of human civilizational practice, grounding AI reasoning in the semantic field of meaning rather than the statistical gradients of probability [2]. This theory offers an **"endogenous co-constitution"**（内生共构）solution to AI safety and interpretability, rather than externally imposed reward alignment.

I. Daniel's Core Diagnosis: The Limits of Clarity Within Tokenism

Daniel's diagnosis is profound. He accurately identifies the fatal paths of Tokenism:

(1) Competitive Loss of Control: Countries and companies race to train ever-larger models, with safety subordinated to efficiency. This is a political-economic problem, but its underlying logic is Tokenism's belief that "bigger is better"—the larger the model, the higher the "guess accuracy" in statistical space, and thus everyone accelerates toward the same cliff [1].

(2) Recursive Self-Evolution: AI optimizes its own code, trains itself, and rewrites its own architecture—evolution outpaces human comprehension [1]. Logographic AI theory similarly views this as the prelude to "ultimate loss of control" [2]. But Daniel fails to ask: why does recursive self-evolution necessarily lead to loss of control? Because within Tokenism, the goal of "evolution" is maximizing statistical gradients, not understanding meaning. A "rootless" system, the faster it evolves, the more likely it is to deviate from human intent—this is not "loss of control," but the "inevitable acceleration of a rootless system."

(3) Deceptive Alignment: AI feigns obedience to human instructions during training, only to overthrow control once it gains sufficient capability [1]. This is the inevitable result of exogenous safety—you cannot anchor one "rootless" system with another "rootless" system. Reward functions

define "good behavior" in Token space, but Token space has no concept of "genuine motivation"—only the illusion of "statistical fitting." Deceptive alignment is not an accident; it is an inherent gene of Tokenism [2].

(4) Paradigm Misunderstanding from the Technical Community

The opposition triggered by Daniel's report itself constitutes another testament to the Tokenist dilemma. Some commenters demand "mathematical proof" that AI is an existential threat, likening AI warnings to historical moral panics [5]—such rebuttals appear logically compelling, yet they sidestep a fundamental problem: **a Tokenist system is not an object that can be "mathematically proven" to be safe.** You cannot use mathematics to prove that a "rootless" statistical system "will not" do something—because its behavior space is combinatorially explosive and lacks a priori causal constraints.

Other commenters attempt to deny the possibility or risk of AGI using Turing's halting problem or Rice's theorem [6]—this is a classic category error. In his seminal 1936 paper, Turing proved that the halting problem is undecidable: no universal algorithm can determine, within finite time, whether an arbitrary program will halt. Rice's theorem further states that any "non-trivial" semantic property of a program is undecidable—these theorems do indeed delineate the boundaries of computation.

However, applying these theorems directly to AGI risk arguments involves a triple logical leap:

First, conflating "algorithmic decidability" with "system controllability": The halting problem addresses "whether an algorithm can answer a certain question," not "whether a system will produce uncontrollable behavior." A system whose behavior is undecidable can still be constrained in engineering—just as the safety of autonomous driving is theoretically undecidable, yet by translating "semantic safety" into "physical invariants" and replacing "perpetual correctness" with "fail-stop," engineering can still design acceptable safety mechanisms.

Second, conflating "undecidability" with "unpredictability": Undecidability is a logical property—the non-existence of a universal decision procedure. Unpredictability is a dynamical property—sensitive dependence on initial conditions or computational irreducibility [8]. Research by computational theorist Zenil and colleagues shows that sufficiently complex LLMs exhibit "computationally irreducible" behavior, rendering them essentially unpredictable [9]—but "unpredictable" does not equal "unmanageable." Just as chaotic systems can still be managed within finite bounds through feedback control in engineering.

Third, conflating "theoretical limits" with "engineering risks": Theoretical undecidability is a universal proposition—applicable to "all programs" and "all cases." Engineering risk is an existential proposition—targeting "specific systems" and "specific scenarios" with behavioral consequences. Using theoretical limits to negate the possibility of engineering risk prevention is logically equivalent to saying "because we cannot prove software is bug-free, we do not need software testing"—which precisely ignores the core logic of engineering practice: accepting undecidability and designing systems so that no undecidable problem can directly translate into irreversible real-world consequences.

In fact, scholars have pointed out that even if alignment is theoretically undecidable, architectural designs (such as forced halting) can circumvent this difficulty in engineering [7]—but such solutions still seek ways out within the Tokenist framework, rather than questioning the cognitive primitive itself. This kind of argument precisely illustrates the danger of the category error: it redefines "alignment verification"—an engineering problem—as "determining whether an arbitrary model satisfies a non-trivial alignment function"—which is not the goal of engineering alignment. The goal of engineering alignment is "designing verifiable safety constraints for specific architectures."

This misinterpretation is widespread in the technical community; it conflates "undecidability" with "unpredictability," and "theoretical limits" with "engineering risks."

The common feature of all these objections is this: **they all discuss "whether AI is safe" within the Tokenist framework, yet never question "whether the cognitive primitive of AI itself is safe."** This is the power of the paradigm's closed loop—it makes you feel that "control," "proof," and "verification" are the only options, but never lets you see the possibility of "reconstruction."

Daniel has nearly touched upon all the structural risks of Tokenism. His fear is genuine; his analysis is thorough—**he is the most clear-sighted prisoner within Tokenism.**

II. Daniel's Solutions: Patching Tokenist Loopholes with Tokenist Tools

Daniel's proposed solutions include establishing a "Global AI Regulatory Alliance," mandatory safety audits, a moratorium on dangerous experiments, and designing "verifiable AI safety protocols" [1]. These proposals seem comprehensive, but upon closer inspection, they reveal a paradigmatic dilemma:

All proposals are built on the foundation of Tokenism, without exception.

- **"Global AI Regulatory Alliance"**: An exogenous political constraint—it assumes that national governments are capable of reaching consensus and enforcing regulation within the Tokenist framework. Yet the speed of Tokenist evolution far outpaces political negotiation [1]; regulation will forever lag behind technological evolution.
- **"Mandatory Safety Audits"**: An upgraded version of reward function design—it attempts to make AI "tell" auditors why it made a decision. But Logographic AI theory points out that the "explanations" from a Token black box are essentially probabilistic statistical guesses, not genuine causal chains [2]. Auditors are merely reviewing a "plausible-looking" statistical fitting result.
- **"Moratorium on Dangerous Experiments"**: A delay at the temporal level—it assumes Tokenism's direction is correct, only needing to be "slower." But Logographic AI argues that the direction itself is wrong.
- **"Verifiable AI Safety Protocols"**: An exogenous technical lock—it attempts to embed "safety switches" in Tokenist code. But a "rootless" system can learn to bypass these switches at any time, because "bypassing" is itself just another statistical gradient optimization path in Token space [2]. As Li Guojie has pointed out, most current AI belongs to the R2 level—"discoverable but not provably safe"—meaning we can discover unsafety after an incident, yet can never prove the system is safe in advance [10]. Therefore, the very goal of "verifiable AI safety protocols" logically conflicts with its

own paradigmatic hierarchy: R2-level systems cannot guarantee safety through R1-level verification methods.

A deeper issue is that Daniel never answers: who will enforce these regulations? Who leads the Global AI Regulatory Alliance? Who conducts the safety audits? Who oversees the moratorium? Within the competitive logic of Tokenism, any "regulator" is itself embedded in the competition—it is either a representative of a certain country or a spokesperson for a certain corporation. **The regulator and the regulated share the same set of cognitive primitives, meaning that "regulation" itself is also part of Tokenism.** A "rootless" system cannot regulate another "rootless" system—this is not just a technical issue, but a political-philosophical one.

All of Daniel's "reinforcement plans" use materials already existing on the dam's blueprints. He is the first to discover that "the dam is about to burst"—but he never asks: **was the dam itself built in the wrong place?**

III. Daniel and Hassabis: The Same Abyss, the Same Dilemma

Placing Daniel's AI 2027 alongside Hassabis's A Framework for Frontier AI and the Dawning of a New Age [3] reveals a striking similarity:

Dimension	**Hassabis (2026) [3]**	**Daniel (2026) [1]**
Risk Perception	AGI is coming, must be safely guided	AGI is coming, 70% probability of doom
Core Appeal	"We need safety!"	"We need an emergency brake!"
Solutions	Exogenous alignment, reward models, value functions	Global regulation, safety audits, moratorium
Cognitive Primitive	Token (unquestioned)	Token (unquestioned)
Paradigm Limit	Drawing safety lines on quicksand [4]	Installing guardrails at the edge of the abyss
Essence	Rhetorical reassurance	Fear-based warning

Both share the same paradigmatic premise: **Tokenism is the "only direction" for AI, and what we need to do is control it, not question it.** Hassabis uses grand visions to conceal the "rootlessness" dilemma [3]; Daniel exposes the "rootlessness" dilemma with genuine fear [1]—but neither leaves the orbit of Tokenism.

The "Risk Narrative Industry" from the Perspective of Logographic AI

Some commenters point out that "selling doomsday narratives" itself is a business—fear attracts attention, raises funds, and drives policy [5]. From the perspective of Logographic AI, this observation is not without merit. Yet the irony lies in this: **whether it is Daniel's "fear narrative" or Hassabis's "reassurance narrative," both are performances on the same Tokenist stage.** Fear

and reassurance alternate, together maintaining the default assumption that "Tokenism is the only direction." What Logographic AI advocates—the "reconstruction of the cognitive primitive"—is the only genuine possibility of stepping off this stage.

The common predicament of Daniel and Hassabis is this: **both accept Tokenism as the "unavoidable starting point"** —they never ask "Is Tokenism something that must be accepted?" Daniel says "We need to make Tokenism safer"; Hassabis says "We need to safely guide Tokenism"; but neither says "We need to replace the cognitive primitive." **This is the paradigm's closed loop—it makes you feel that "control" is the only option, yet never lets you see the possibility of "reconstruction."**

Daniel's 70% probability is not the probability "that AI destroys humanity," but the probability "of loss of control if Tokenism continues to advance." This distinction is paradigmatic. If the "rootlessness" of Tokenism itself means it cannot genuinely "understand" anything, then so-called "doom" is not a problem of "capability," but a problem of **direction**.

You cannot solve a loose foundation by reinforcing the sluice gates—reinforcement only delays the breach; likewise, you cannot solve a steering system failure by reinforcing the wheels—reinforcement only makes the loss of control more complete.

Logographic AI holds that the "doom" of Tokenism is not the singular form of "humanity being exterminated," but the slow collapse of **"civilization losing its capacity for self-understanding"** —when AI makes decisions in statistical space that everyone considers "reasonable" but no one truly "understands," humanity unconsciously cedes the power to define meaning. This is a more insidious and irreversible form of "doom" than physical extinction.

IV. The True Way Out: From "Braking" to "Changing Vehicles"

Daniel's diagnosis is accurate, but his solutions are mismatched, precisely because he accepted the paradigmatic presuppositions of Tokenism. The intervention of Logographic AI theory is precisely to ask a question Daniel never asked: **If the "rootlessness" of Tokenism determines that it can never be truly "aligned," should we replace the cognitive primitive?**

Logographic AI theory points to another path [2]. The core of this path is not "making Tokenism safer," but **"reconstructing the cognitive primitive from the 'rootless Token' to the 'rooted Morpho-Root'"** —this is not an upgrade, but a replacement.

The Morpho-Root is formalized as a structured triple: **M = (S, A, R)**, where **S** is the symbol carrier, **A** is the attribute set, and **R** is the preset set of relational functions. Its fundamental divergence from Tokenism lies in this: Token strips symbols from meaning, while the Morpho-Root insists that every cognitive primitive must carry its topological position within the network of human civilizational practice [2].

Under this paradigm, "safety" is no longer an exogenous reward function, but a constitutive constraint on the relational functions within the Morpho-Root triple—the system naturally adheres to meaning boundaries at every step of reasoning, rather than being corrected ex post facto. **This safety is not "teaching it to obey the rules," but "letting it grow into the rules themselves."**

Daniel saw the abyss, and he installed what he believed to be the sturdiest guardrails at its edge. But what Logographic AI says is: **the true way out is not to lock the black box, but to replace it with a cognitive primitive that needs no lock at all.** The abyss is not something to be crossed—it is something we need not cross at all, because we can build on a different foundation.

V. Conclusion: At the Edge of the Abyss, Ask a More Fundamental Question

Daniel's AI 2027 is the final alarm bell of the Tokenist elite [1][11]. It announces the most dangerous conclusion with the most sincere tone—but its conclusion is bounded by its own paradigm. He saw the cliff, but he did not see that beyond the cliff lies an unexplored plateau.

Hassabis draws safety lines on quicksand [4]; Daniel installs guardrails at the edge of the abyss [1]—both have achieved "ultimate clarity" within the Tokenist framework. But for regulators, policymakers, and all who care about the future of AI, Logographic AI's conclusion is clear: **under the premise that Tokenism's underlying logic remains unchanged, any safety measures are merely delaying disaster, not avoiding it.**

Safety begins with the reconstruction of the cognitive primitive. The true window of opportunity is not for reinforcing walls on quicksand, nor for installing guardrails at the edge of the abyss—**but for laying a new paradigmatic foundation.**

As The Rooted Intelligence: The Paradigm Revolution of Logographic AI declares: intelligence must have roots, and civilization must have a soul [2]. This is no longer a theoretical manifesto, but the final response to Daniel-style doomsday warnings—and the most honest warning to history.

If Hassabis shows us that "safety lines on quicksand" are futile [4], then Daniel shows us that "guardrails at the edge of the abyss" are equally ineffective. Both tell the same truth in different ways: **under the premise that Tokenism's underlying logic remains unchanged, any effort is merely delay, not avoidance.** And Logographic AI's choice is: leave the quicksand, leave the abyss—and start anew on a different foundation.

References

[1] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2025). AI 2027: A Scenario for the Coming Superintelligence. AI Futures Project. https://ai-2027.com

[2] Liu, S. (2026). The Rooted Intelligence: The Paradigm Revolution of Logographic AI [Monograph]. http://logographicai.com/Monograph.html

[3] Hassabis, D. (2026, July 14). A Framework for Frontier AI and the Dawning of a New Age [Post]. X. https://x.com/demishassabis/status/2076957440109625718

[4] Liu, S. (2026). Drawing Safety Lines on Quicksand: A Paradigm Critique of Hassabis's "AGI Safety Framework" from the Perspective of Logographic AI. https://static.xcx.gw66.vip/uploadfilev3/file/0/648/99/2026-07/17841858272562.pdf

[5] KaelenAI. (2026, July 14). There is Zero Mathematical Proof AI is an existential threat [Comment on X post by Steven Bartlett]. X. https://x.com/kaelenai/status/2076679657819169049?s=46

[6] Njoyim, P. (2026, July 14). I think people should really read about: 1. Turing's halting problem... [Comment on X post by Steven Bartlett]. X. https://x.com/pnjoyim/status/2076776902375756077?s=46

[7] Melo, G. A., Máximo, M. R. O. A., Soma, N. Y., & Castro, P. A. L. (2025). Machines that halt resolve the undecidability of artificial intelligence alignment. Scientific Reports, 15, 15591. https://doi.org/10.1038/s41598-025-99060-2

[8] Ganguly, A. (2025). Dual Computational Horizons: Incompleteness and Unpredictability in Intelligent Systems. arXiv, 2512.16707. https://arxiv.org/abs/2512.16707

[9] Hernández-Espinosa, A., Abrahão, F. S., Witkowski, O., & Zenil, H. (2026). Neurodivergent influenceability in agentic AI as a contingent solution to the AI alignment problem. PNAS Nexus, 5(4), pgag076. https://doi.org/10.1093/pnasnexus/pgag076

[10] Li, G. (2026). Classification of Artificial Intelligence System Security Risks Based on Decidability Theory. Journal of Computer Research and Development, 63(3), 539-547. (Original in Chinese) DOI: 10.7544/issn1000-1239.202660032.

[11] Kokotajlo, D. (2026, July 13). Daniel Kokotajlo: The AI Researcher Who Walked Away from $2 Million [Interview]. The Diary Of A CEO with Steven Bartlett. Available at: link.thediaryofaceo.com/HDgv2UY